COMP 605: Graduate Seminar in Learning Theory

Lecture 1

Maryam Aliakbarpour

Fall 2024

Today's lecture

- Introduction
- Class format
- Policies
- Introduction to the topic

Introduction

Instructor: Maryam Aliakbarpour

Email: maryama@rice.edu

Office hour: By appointment (email me)

Lectures: Wednesdays 4-5:15pm, Duncan Hall 1075

Website: https://maryamaliakbarpour.com/courses/F24/index.html+ Canvas

Please turn on your notification on Canvas!

Class objectives

Studying fundamental problems in learning theory from a new perspective:

- Computational aspects: limited time or memory
- Societal aspects: privacy and fairness

We will return to this!

Practicing research soft skills:

- How to approach a problem
- How to review / write a paper
- Presenting technical material

Class Prerequisites

- solid understanding of mathematical proofs
- basic algorithms, and probability
- A graduate level course in algorithms or machine learning is recommended.

Class format

- In each class, we focus on one topic / one paper.

- Before class:
  - Reading assignment: read the paper
  - Provide a review on canvas

- Presentation:
  - A student presents a topic or a paper (1hr presentation)

- Questions / Discussion

Class format

- A list of suggested papers: → Syllabus
- You may also pick papers that are not listed but are relevant to the topic of the class.
- Sign up for your presentation here, and fill out this form by Thursday (9/12).

Class format: presentation

A 1-hr long presentation:

- Introduction: What and why?
- Related work
- Problem definition
- Solution
- Technical part*

Class format: presentation

Practice your talk! (many times)

(Optional) Meet with me on Monday or Tuesday before your presentation.

- Set an appointment (maryama@rice.edu)

Class format: reading assignment

Read the paper before class, and be present.

Think of it as a mini-review.

Canvas assignment:

- Summary of the paper.
- Your opinion: Strengths / Limitations. Next steps?

Class format: class project

Only if you register for 3-credit

Two options:

- Survey of results
- Research project

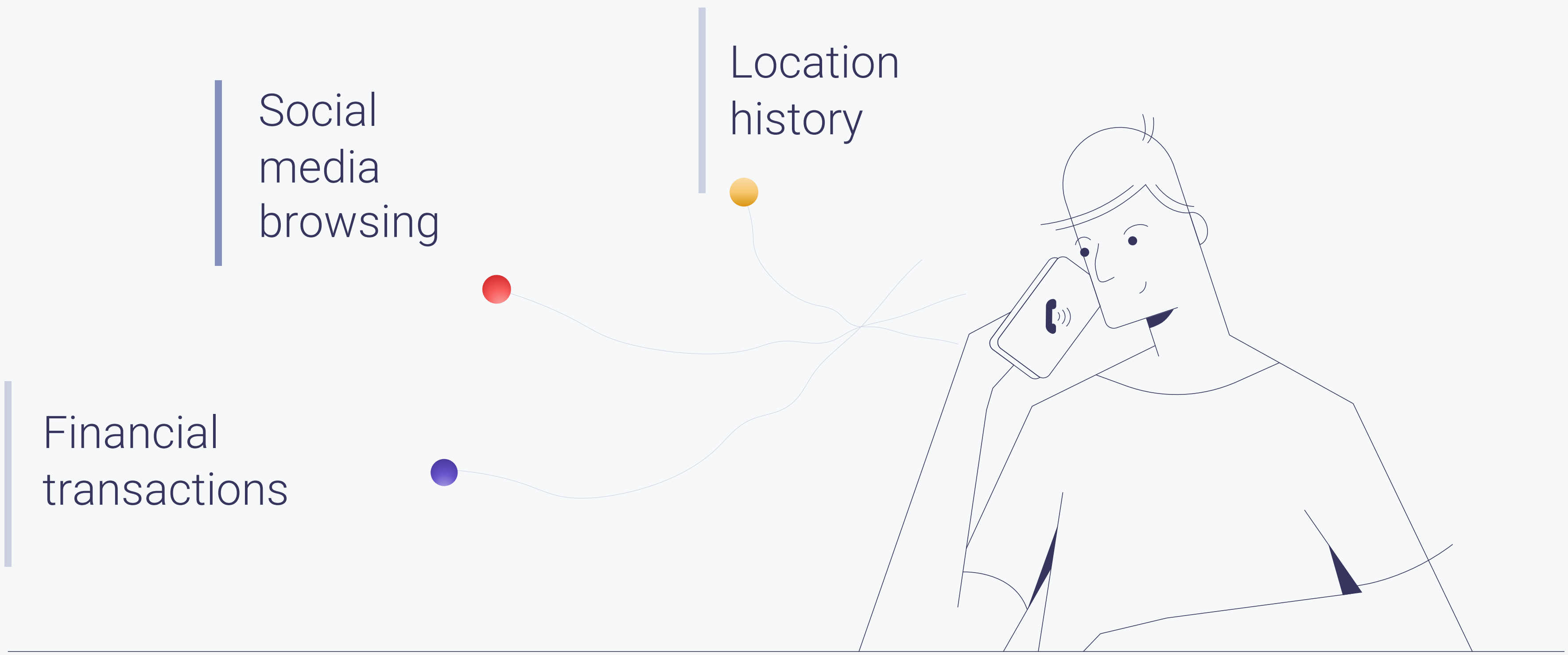
Policies

Read Syllabus

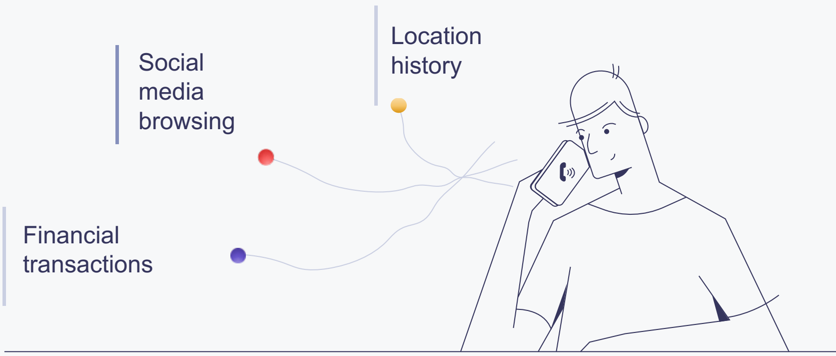
- An inclusive environment
- Rice Honor Code
- Disability Resource Center
- Wellbeing and Mental Health
- Title IX Responsible Employee Notification

Our topic

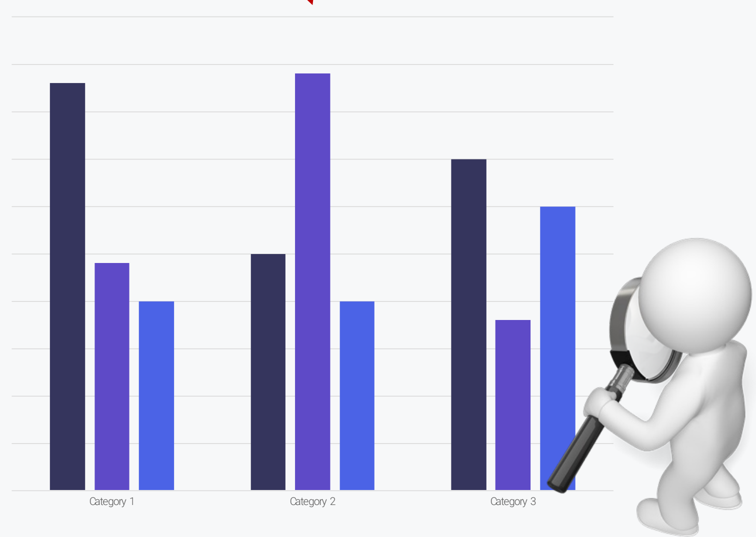
Our daily activities produce vast amounts of data.



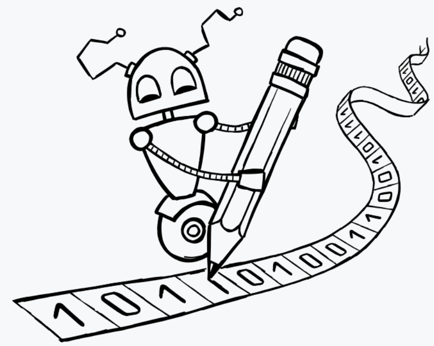
Our daily activities produce vast amounts of data.



How can we extract meaningful information?



Statistical inference



Data:
samples from D
$x_1, x_2, \dots, x_m$

$\longrightarrow$ Algorithm $\longrightarrow$

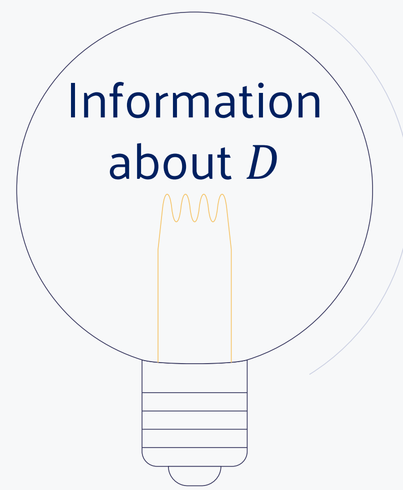
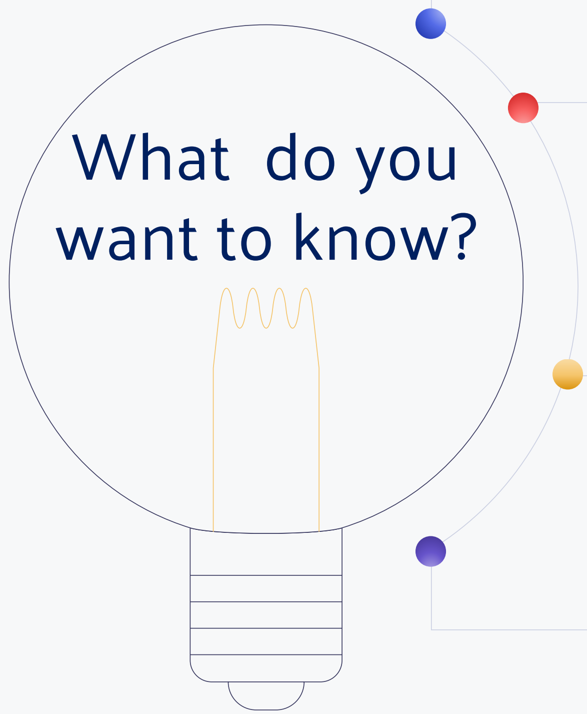


Image from: https://tilics.dmi.unibas.ch/the-turing-machine

Statistical inference

Data:
samples from D
$x_1, x_2, \ldots, x_m$



Estimation:
Estimate parameters of distribution
e.g. mean, variance

Testing:
Test distribution D has a specific property
e.g. uniformity, unimodal

Learning:
Learn distribution D in a class
e.g. Gaussians

Classification:
Learn a classifier from labeled data
e.g. learning half-spaces

Classical goal: data efficiency

New goal: understanding the relationship between all of these aspects

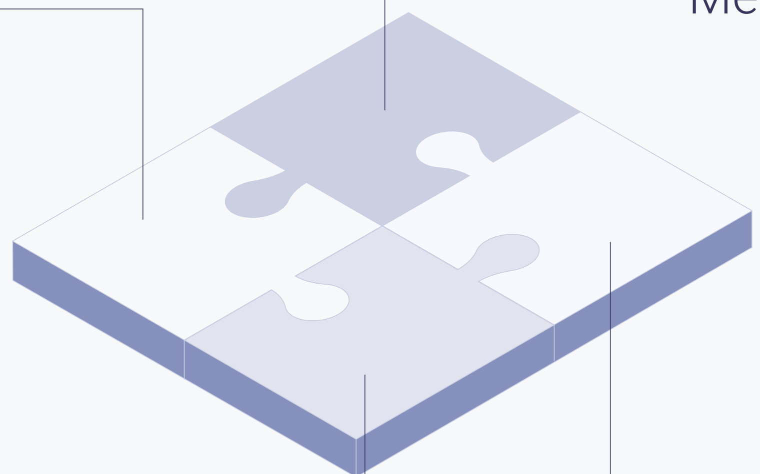
Use as few data points as possible

Computational aspect:

Memory / time

Accuracy

Dependencies on the error parameter



Sample complexity

data points

Societal aspect:

Privacy / Fairness

Statistical inference

Data:
samples from D
$x_1, x_2, \dots, x_m$

Algorithm with
limited memory
limited time
private
fair

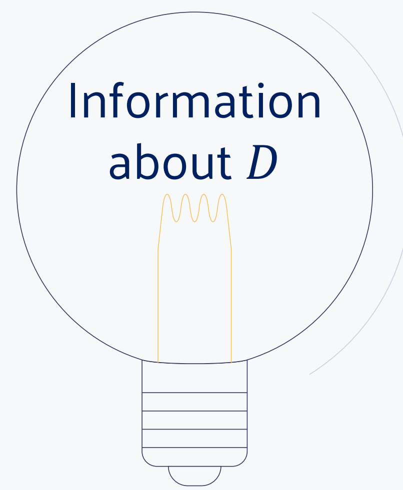
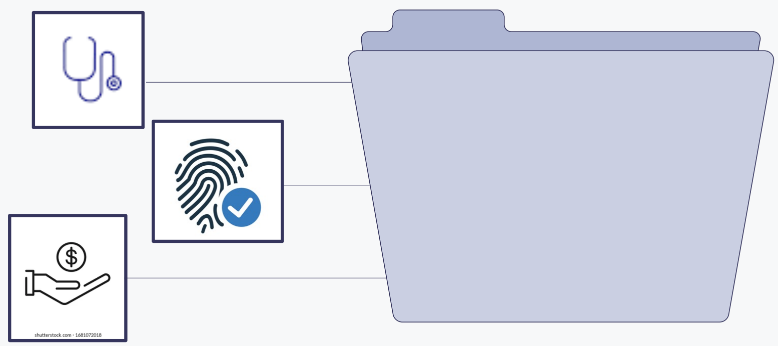


Image from: https://tilics.dmi.unibas.ch/the-turing-machine

This talk

Part I: Inference with privacy

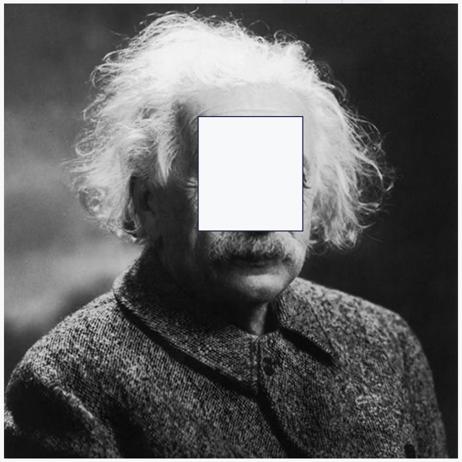
Part II: Inference with limited memory



Sensitive data requires privacy preserving algorithms.

Privacy

- Learn about community, but not individuals



- Anonymization $\neq$ not-identifiable

  re-identification of Massachusetts Governor's medical data
  within an insurance data set

- Global information leaks information about individuals!

  Example: Average net worth of patients in oncology

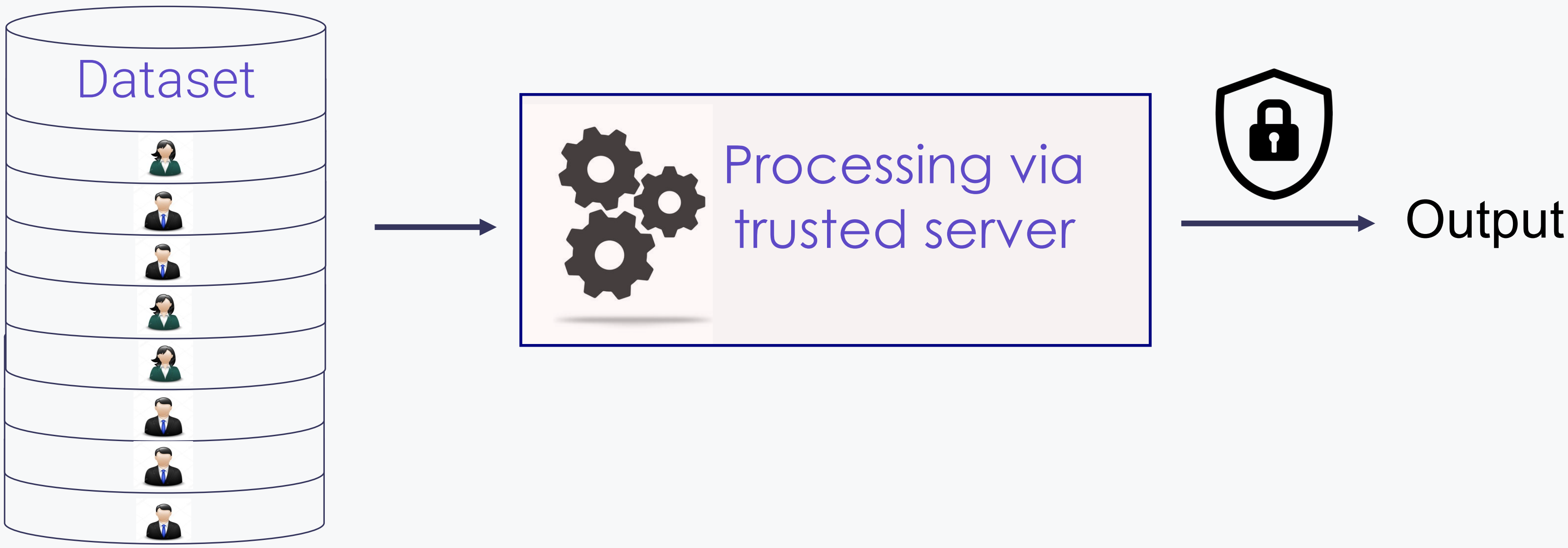
Differential privacy

- Mathematical formulation
- Not ambiguous

  Irrefutable claims
- Extensive use in **practice:**

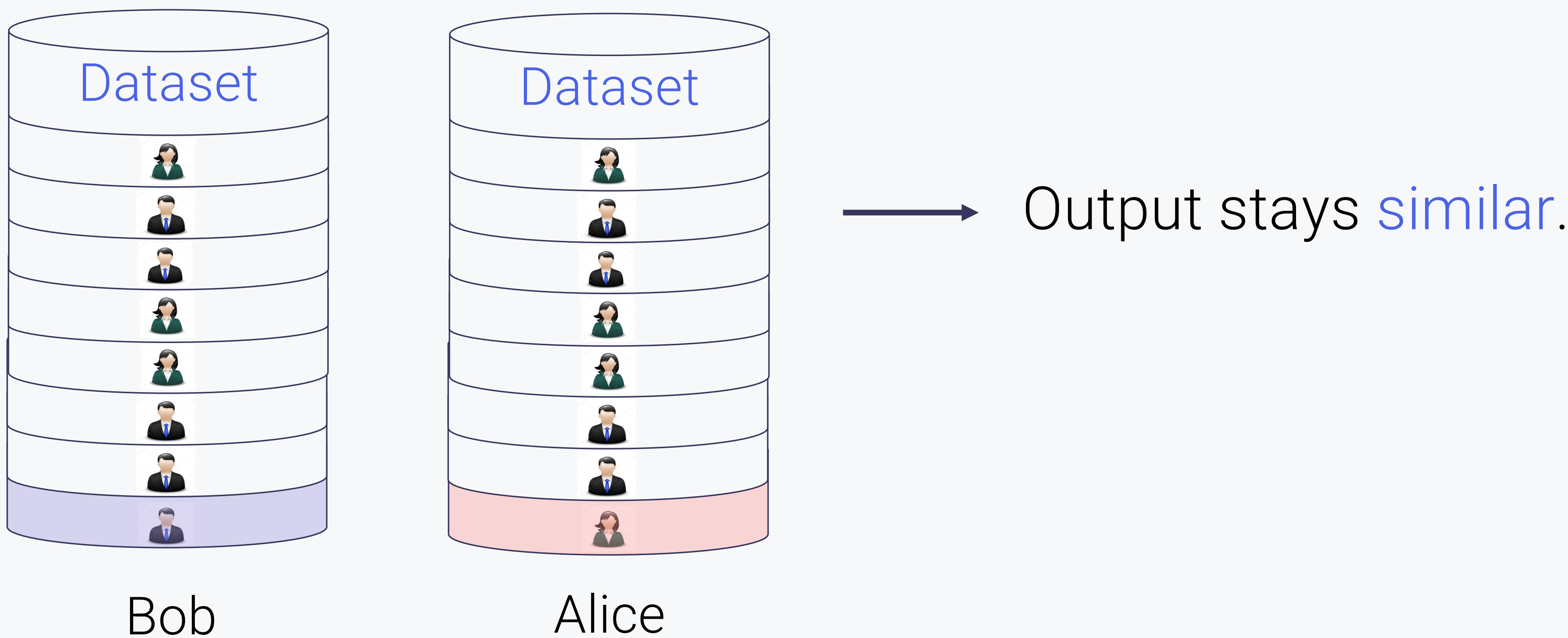
  Apple, Google, US census

Differential privacy (central)



Differential privacy

Output should not depend on a single data point.

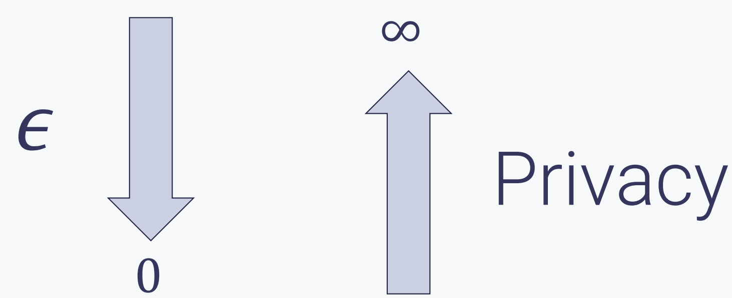


Differential privacy

ϵ-differentially private algorithm A:

- Any possible output Y
- Two neighboring datasets X, X' s.t. they differ in one sample

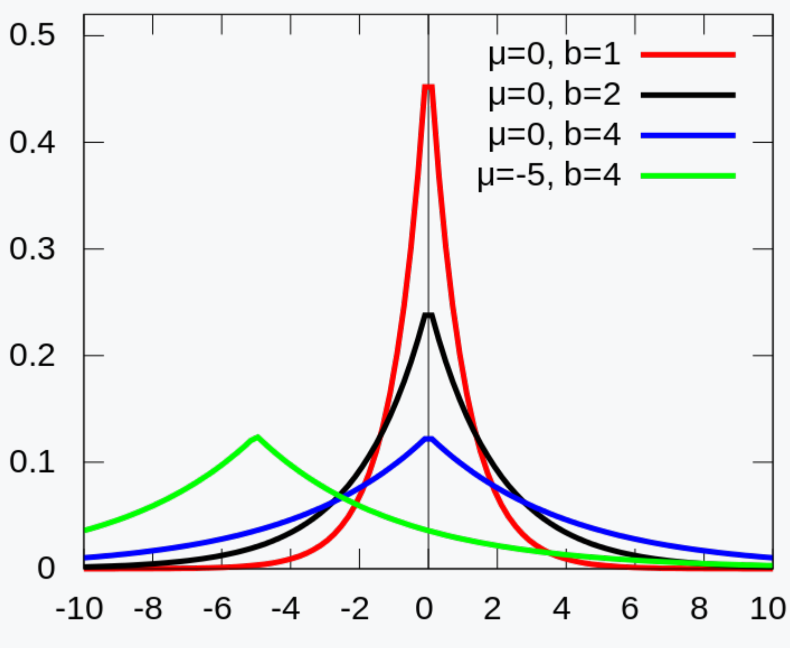
$$\Pr[A(X) = Y] \leq e^{\epsilon} \Pr[A(X') = Y]$$



[Dinur and Nissim'03, Dwork, McSherry, Nissim, and Smith'06, Dwork'06]

Laplace Mechanisms

Laplace distribution



- PDF at point x: $\frac{1}{2\mathrm{b}}\mathrm{exp}\left(-\frac{|x-\mu|}{b}\right)$
- Expected value: 0
- Variance: $2b^2$
- CDF: If $Y{\sim}Lap(b)$ then

$$\Pr[|Y| \geq t] = e^{-t/b}$$

ℓ_1-sensitivity

For two neighboring datasets X, X' such that $|X - X'| = 1$,

the sensitivity of f is:

$$\Delta f \triangleq \max_{X,X'} |f(X) - f(X')|$$

Laplace Mechanism

Can make f a ϵ-differentially private function by adding Laplace noise to it.

Laplace noise

$$\text{Function } f(X) \xrightarrow{+\,\mathrm{Lap}(\Delta f/\epsilon)} \text{Private value } \tilde{f}(X)$$

Usage

Works really well when the sensitivity is small (small noise):

- Count queries
- Histograms
- Low sensitivity statistics: #unseen

Provable guarantees

Theorem: Laplace mechanism is ϵ-differentially private.

Theorem: Laplace mechanism is accurate. For all $\delta \in (0, 1]$:

$$\Pr\left[\left|f(x) - \tilde{f}(x)\right| \geq \frac{\ln\left(\frac{1}{\delta}\right)\Delta f}{\epsilon}\right] \leq \delta$$

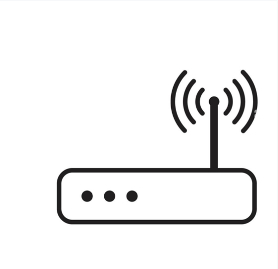
This talk

Part I: Inference with privacy

Part II: Inference with limited memory

Why limited memory?

- Size of working memory < size of data
- Facilitates communication and processing of distributed data
- Insightful: what summarizes the data



Memory restriction can affect learning drastically!

- [Raz, FOCS. 2016]
  Parity learning problem
- [Chien, Ligett, McGregor. ITCS 2010]
  Robust statistics and distribution testing
- [Diakonikolas, Gouleakis, Kane, Rao. COLT 2019]
  Distribution testing
- [Sharam, Sidford, Valiant. STOC 2019]
  Memory-Sample Tradeoffs for Linear Regression
- [Brown, Bun, Smith. COLT 2022]
  Memory lower bounds for sparse linear predictors

And many more…

Memory restriction can affect learning drastically!

[Raz'16]: Fast learning requires good memory!

Parity learning problem:
- Goal: find $w \in \{0,1\}^n$
- Samples: a random $x \in \{0,1\}^n$ and $w \cdot x$

By Gaussian elimination

$O(n^2)$ bits of memory

$O(n)$ samples

[Raz'16]: **Any algorithm using**

$\leq \frac{n^2}{25}$ bits of memory

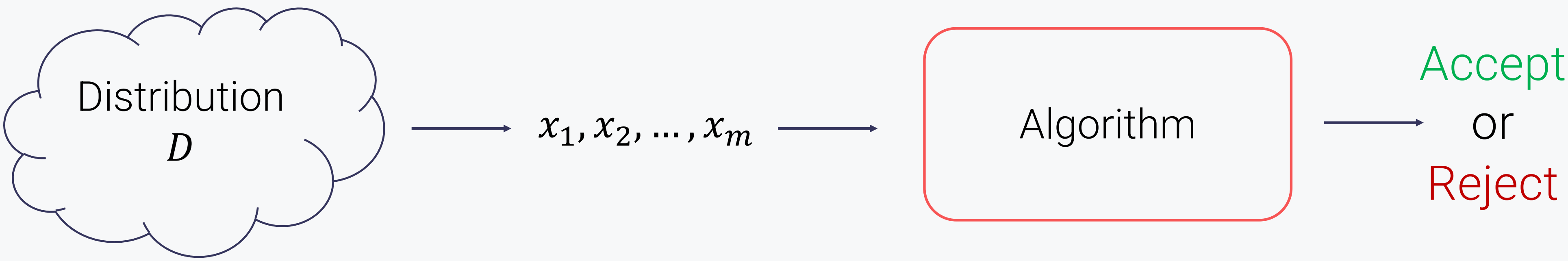
needs exponentially many samples

Example I:
Private Hypothesis Testing

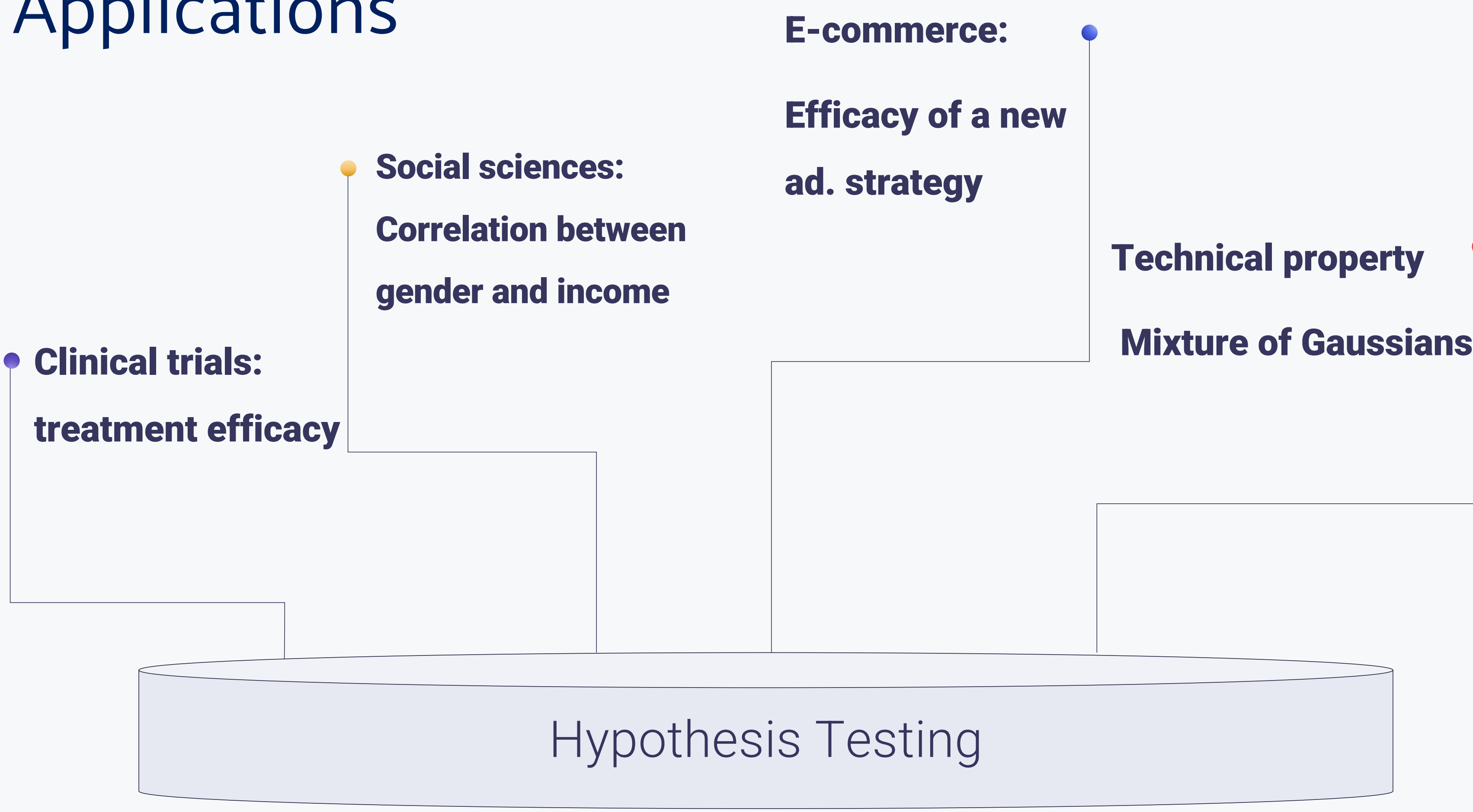
Joint work with Daniel Kane (UCSD), Ilias Diakonikolas (UW Madison), Ronitt Rubinfeld (MIT)

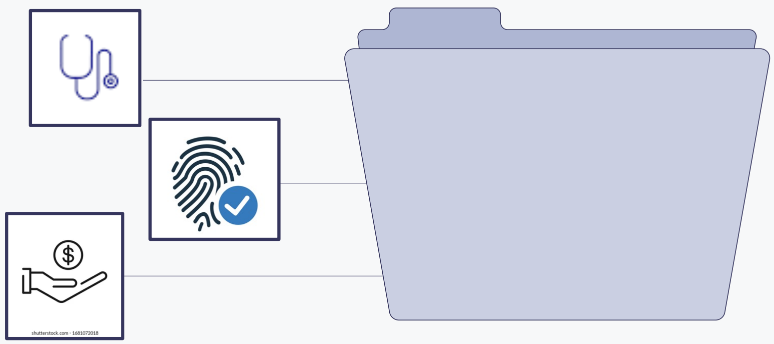
Hypothesis testing

Does D have a particular property or not?



Applications





Sensitive data requires privacy preserving algorithms.

Goal:

Design testing algorithms:

- Accurate
- Optimal number of data points
- Privacy preserving

Active area of research: [Rogers, Roth, Smith, Thakkar'16], [Gaboardi, Lim, Rogers, Vadhan'16], [Cai, Daskalakis, Kamath'17], [A, Diakonikolas, Rubinfeld'18], [Acharya, Sun, Zhang'18]: [Bun, Kamath, Steinke, Wu'19], [Canonne, Kamath, McMillan, Smith, Ullman'19], [Canonne, Kamath, McMillan, Ullman, Zakynthinou'20], [Vepakomma, Amiri, Canonne, Raskar, Pentland'22]

Our problem:

Closeness testing:

Are two distributions equal?

Example: treatment efficacy

Closeness testing:

Are two distributions equal?

Pain level after treatment: 2, 10, 3, 1, 2, 9, 3, 1

Pain level in the control group: 6, 2, 7, 2, 3, 6, 2, 3

Example: treatment efficacy

Closeness testing:

Are two distributions equal?

Number of sold items per day: 2, 10, 3, 1, 2, 9, 3, 1

Number of sold items after price drop: 6, 2, 7, 2, 3, 6, 2, 3

Our problem: closeness testing

Distribution p over $[n]$ → iid samples: $x_1, x_2, \dots, x_s$ →

Distribution q over $[n]$ → iid samples: $y_1, y_2, \dots, y_s$ →

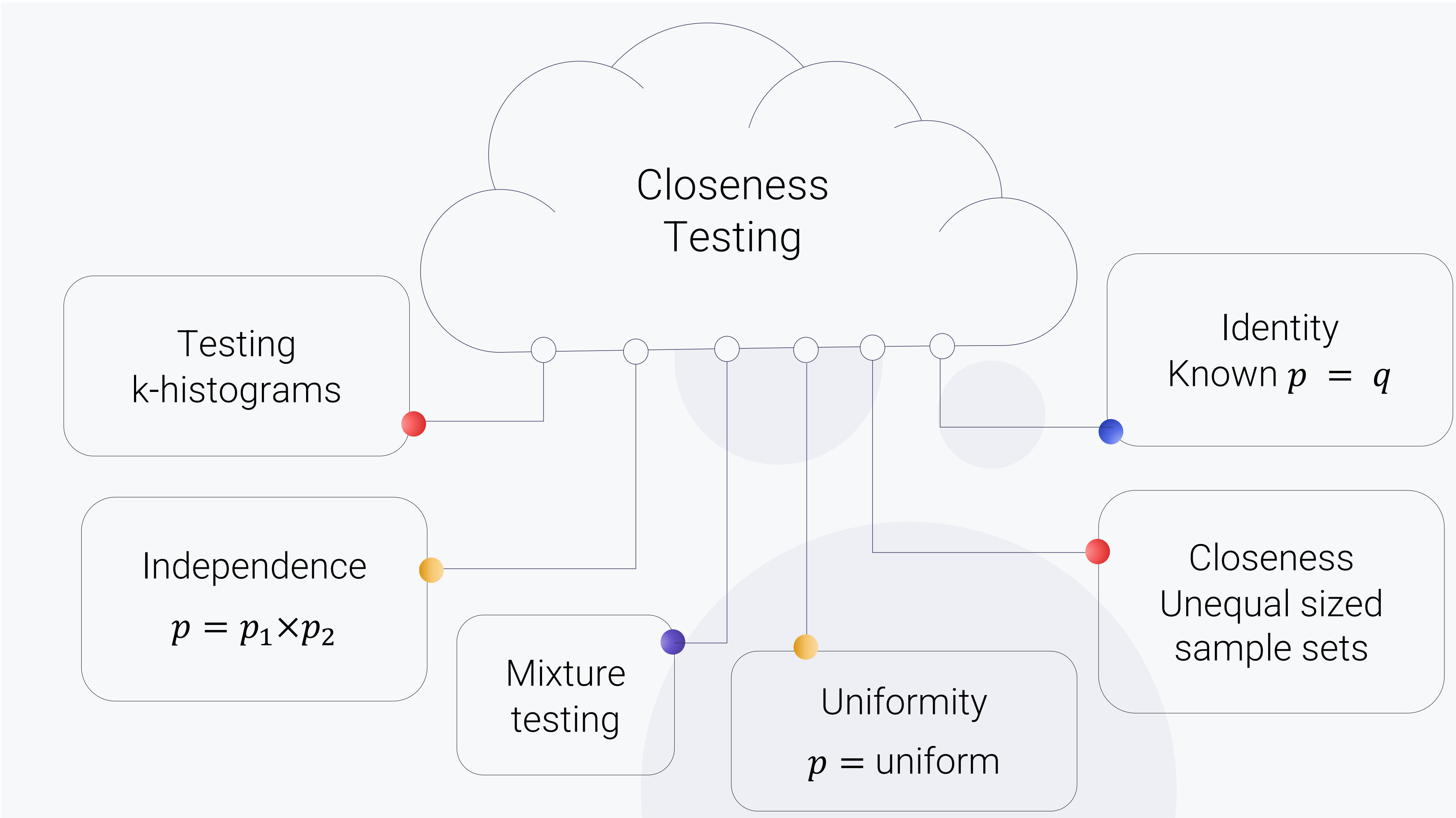
Tester

with prob. 0.9

Output = Accept if $q = p$

Reject if p and q are α-far in ℓ_1-distance

[Batu, Fortnow, Rubinfeld, Smith, White'00]



Closeness Testing
Testing k-histograms
Identity Known $p = q$
Independence $p = p_1 \times p_2$
Closeness Unequal sized sample sets
Mixture testing
Uniformity $p =$ uniform

Closeness testing implies independence testing

$(X, Y) \sim p.$

Question: Are X and Y independent?

p_1 and p_1 are the marginals

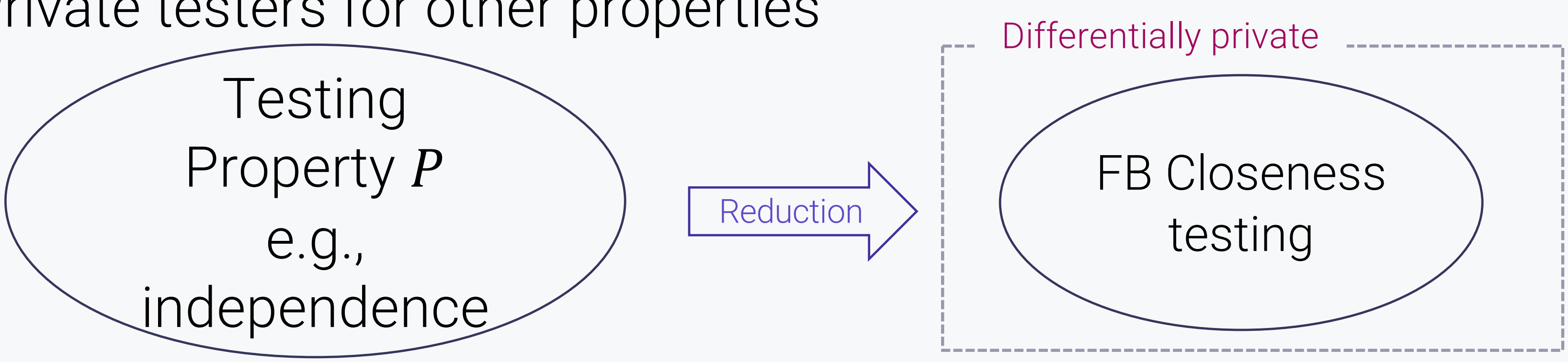
X and Y are independent $\iff$ $p = p_1 \times p_2$

X and Y are far from being independent $\iff$ $|p - p_1 \times p_2|_1 \geq \Theta(\alpha)$

[Batu, Fischer, Fortnow, Kumar, Rubinfeld, White'01]

Our results

- New flattening-based (FB) private tester for closeness testing
- Characterizing the non-private reductions that results in private testers automatically
- Private testers for other properties



[A, Diakonikolas, Kane, Rubinfeld NeurIPS19]

Non-private tester by [Diakonikolas, Kane’16]

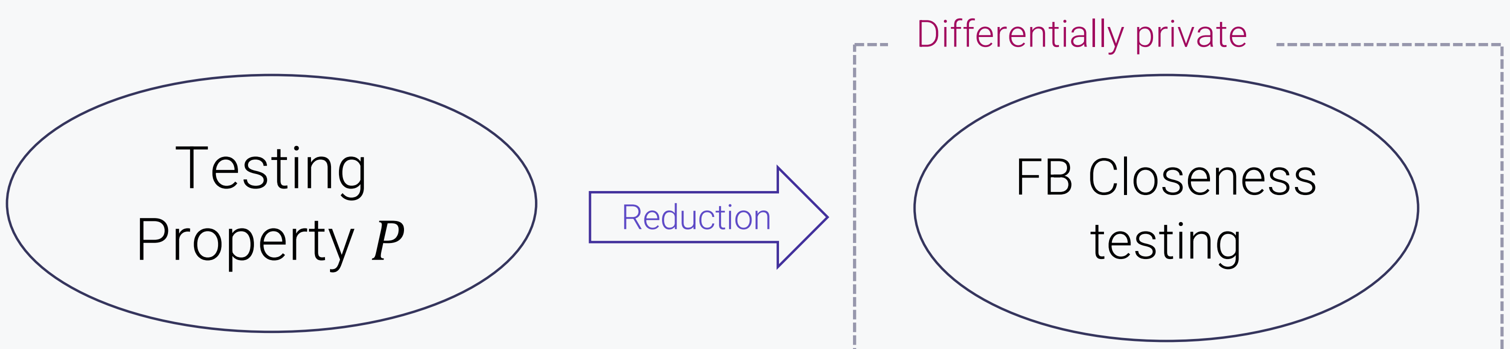
Our results

New flattening-based (FB) private tester

Why this tester?

- Exploits the underlying structure of distributions

- Only known optimal results for some problems



[A, Diakonikolas, Kane, Rubinfeld NeurIPS19]

Our result on closeness: privacy is almost free!

Theorem

[A, Diakonikolas, Kane, Rubinfeld'19]

There exists a ϵ-private algorithm for testing closeness of two distributions p and q over domain of $[n]$ with error parameter α that uses

$$O\Bigg(\underbrace{\frac{n^{2/3}}{\alpha^{4/3}} + \frac{\sqrt{n}}{\alpha^2}}_{\text{Non-private cost}} + \underbrace{\frac{\sqrt{n}}{\alpha\sqrt{\epsilon}} + \frac{1}{\alpha^2 \epsilon}}_{\text{Cost of privacy}}\Bigg)$$

samples from p and q.

Our results on other properties

- New ϵ-DP tester for independence (domain = $[\boldsymbol{n}]\times[\boldsymbol{m}]$ when $m \le n$)

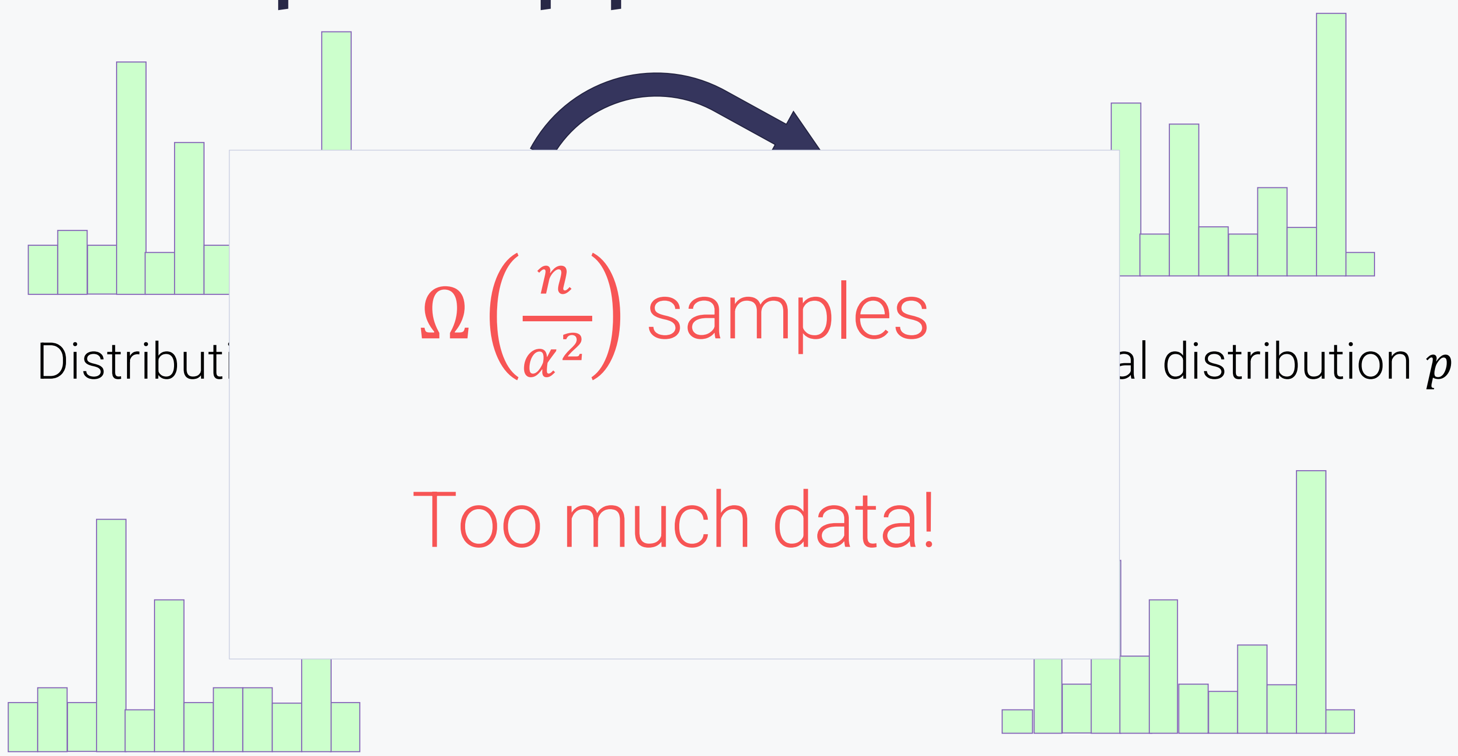
$$O(\underbrace{\boldsymbol{n}^{2/3}\,\boldsymbol{m}^{1/3}/\boldsymbol{\alpha}^{4/3} + \sqrt{\boldsymbol{n}\,\boldsymbol{m}}/\boldsymbol{\alpha}^{2}}_{\text{Non-private cost}} + \underbrace{\sqrt{\boldsymbol{n}\,\boldsymbol{m}\log\boldsymbol{n}}\,/(\boldsymbol{\alpha\epsilon}) + \mathbf{1}/(\boldsymbol{\alpha}^{\mathbf{2}}\boldsymbol{\epsilon})}_{\text{Cost of privacy}})$$

- New ϵ-DP tester for testing closeness with unequal sized samples

- Tighter result for closeness/uniformity/identity

Techniques

How? Simple approach

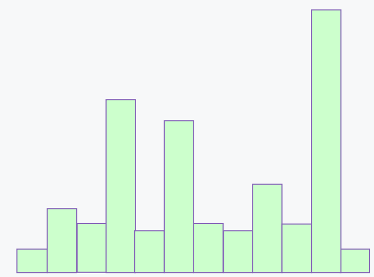


$\Omega\left(\frac{n}{\alpha^2}\right)$ samples

Too much data!

Sub-linear?

Frequency of element i in the sample set = X_i



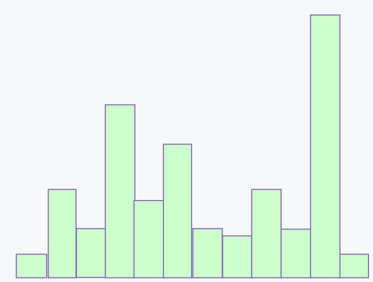
Empirical distribution p

An alternative way:

Statistic $Z \coloneqq \sum_{i=1}^{n}(X_i - Y_i)^2 \; - X_i - Y_i$

$p = q \quad \longleftrightarrow$ Small Z

$|p - q|_1 \geq \alpha \quad \longleftrightarrow$ Large Z



Empirical distribution q

Frequency of element i in the sample set = Y_i

Sub-linear? Potential solution

Statistic: $Z \coloneqq \sum_{i=1}^{n} (X_i - Y_i)^2 \ - X_i - Y_i$

Sample complexity = $\Omega\left(\frac{n \cdot \max(|p|_2, |q|_2)}{\alpha^2}\right)$ $\propto$ max ℓ_2-norm of p and q

$p \rightarrow p'$

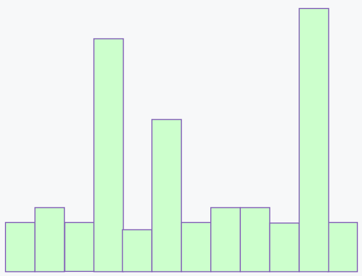
$q \rightarrow q'$

low ℓ_2-norm

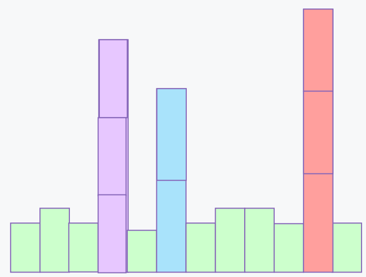
No change in ℓ_1-distance!

$$\begin{cases} p = q \implies p' = q' \\ |p - q|_1 \geq \alpha \implies |p' - q'|_1 \geq \alpha \end{cases}$$

How flattening reduces ℓ_2-norm

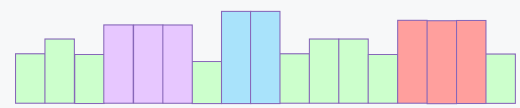


Distribution p



Detecting large elements

On a new domain



Distribution p'

How? Draw samples and see frequencies

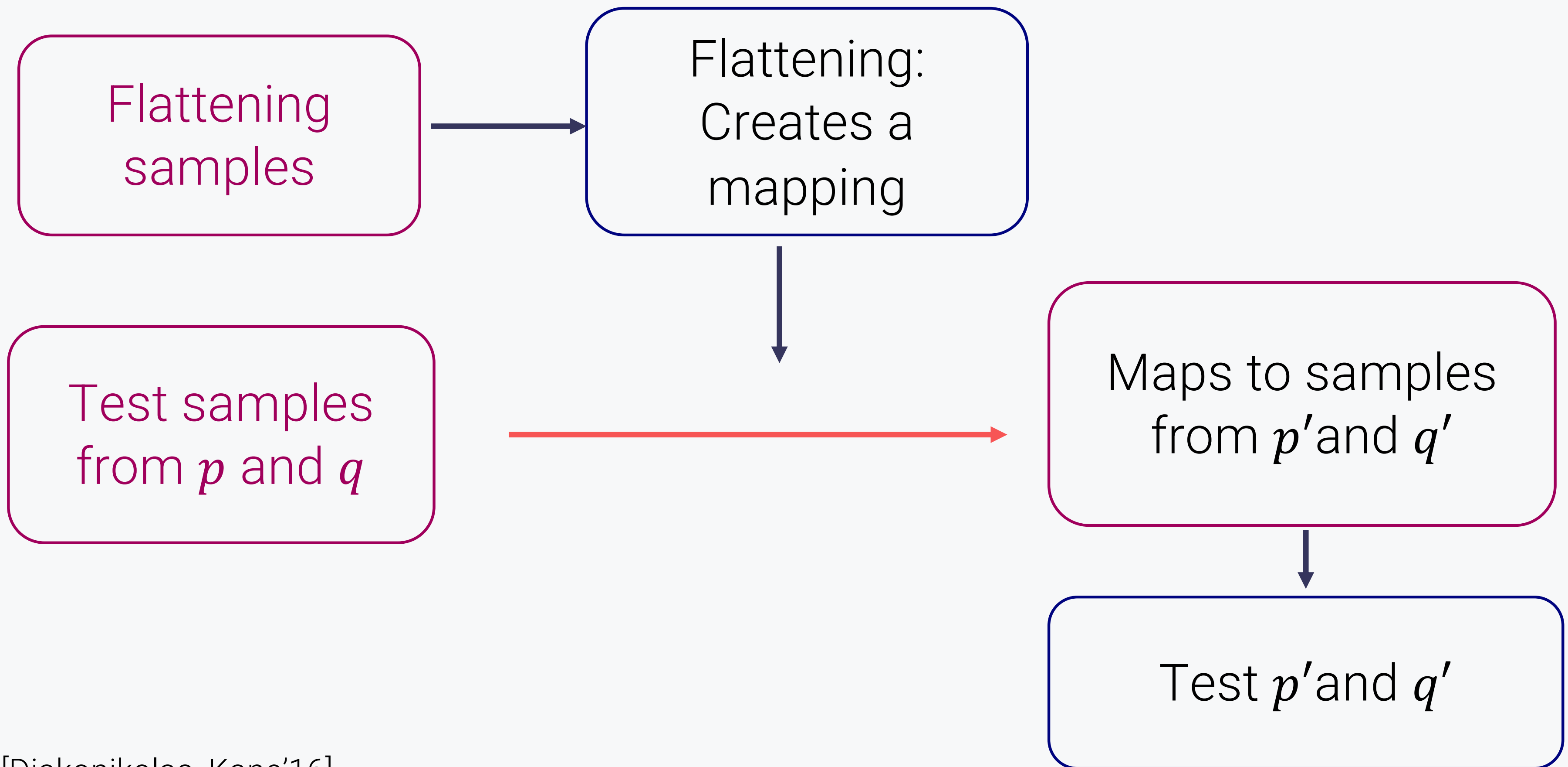
$$E[|p'|_2^2] < \frac{1}{|F|}$$

Flattening Samples F: 

bins = frequency in $F + 1$

[Diakonikolas, Kane'16]

Testing closeness via flattening

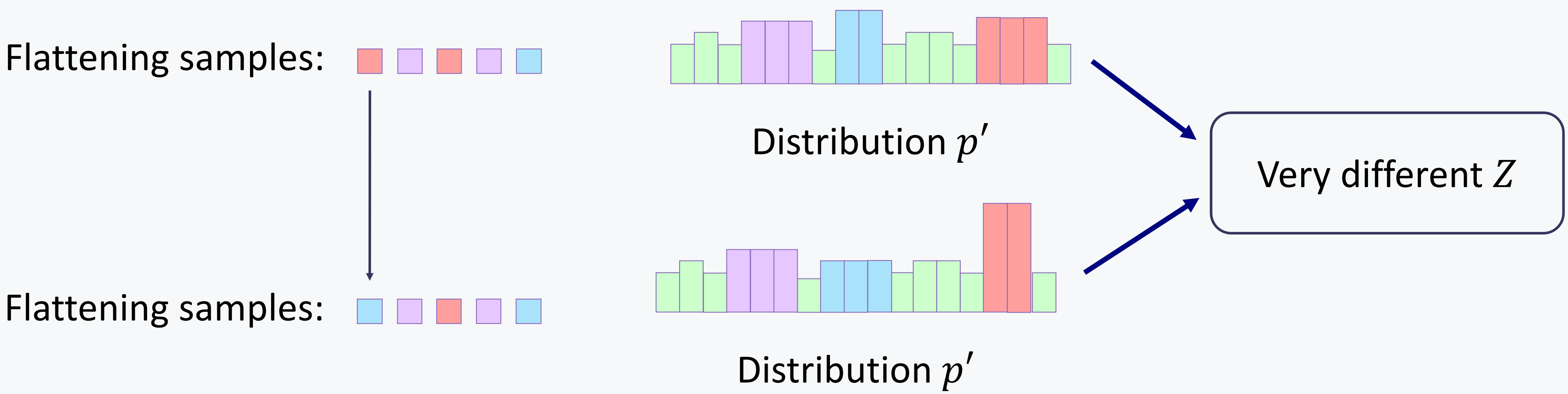


[Diakonikolas, Kane'16]

Not easy to privatize

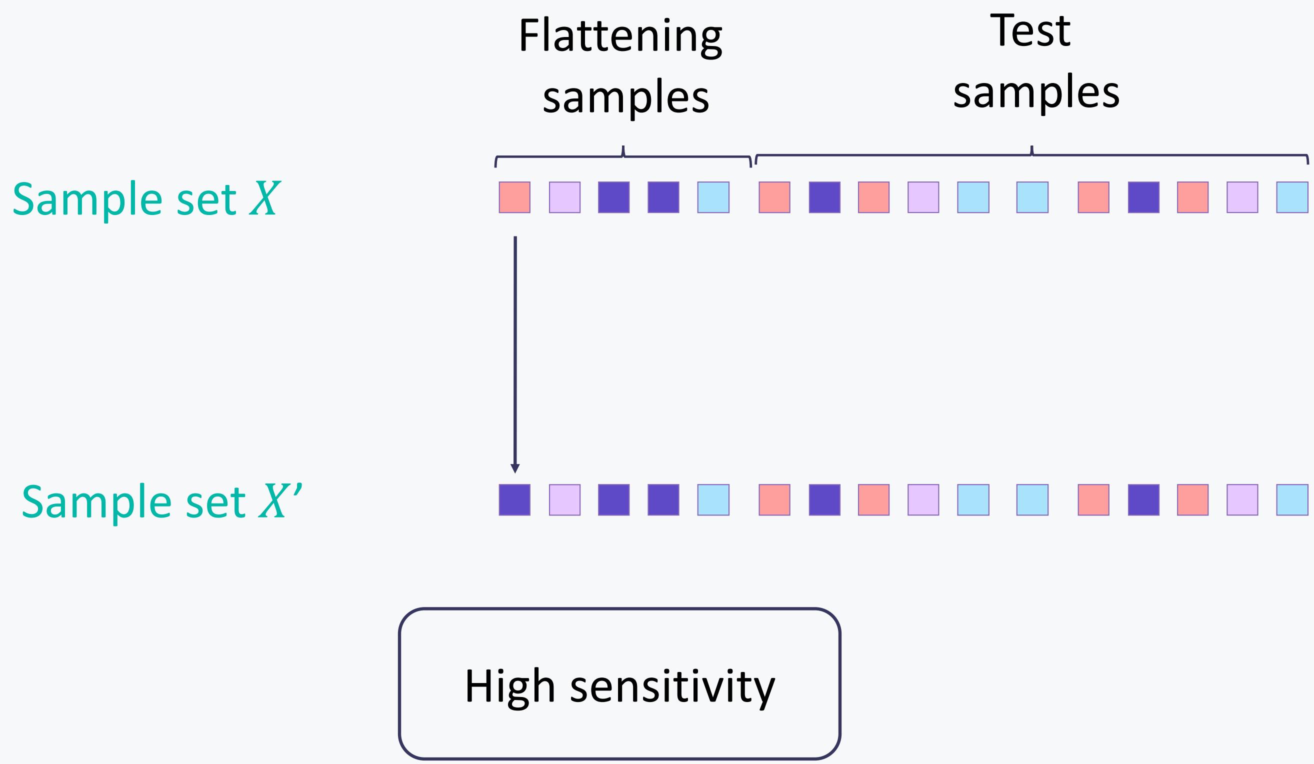
Flattening technique: strong, but sensitive…

Hard to make it private!

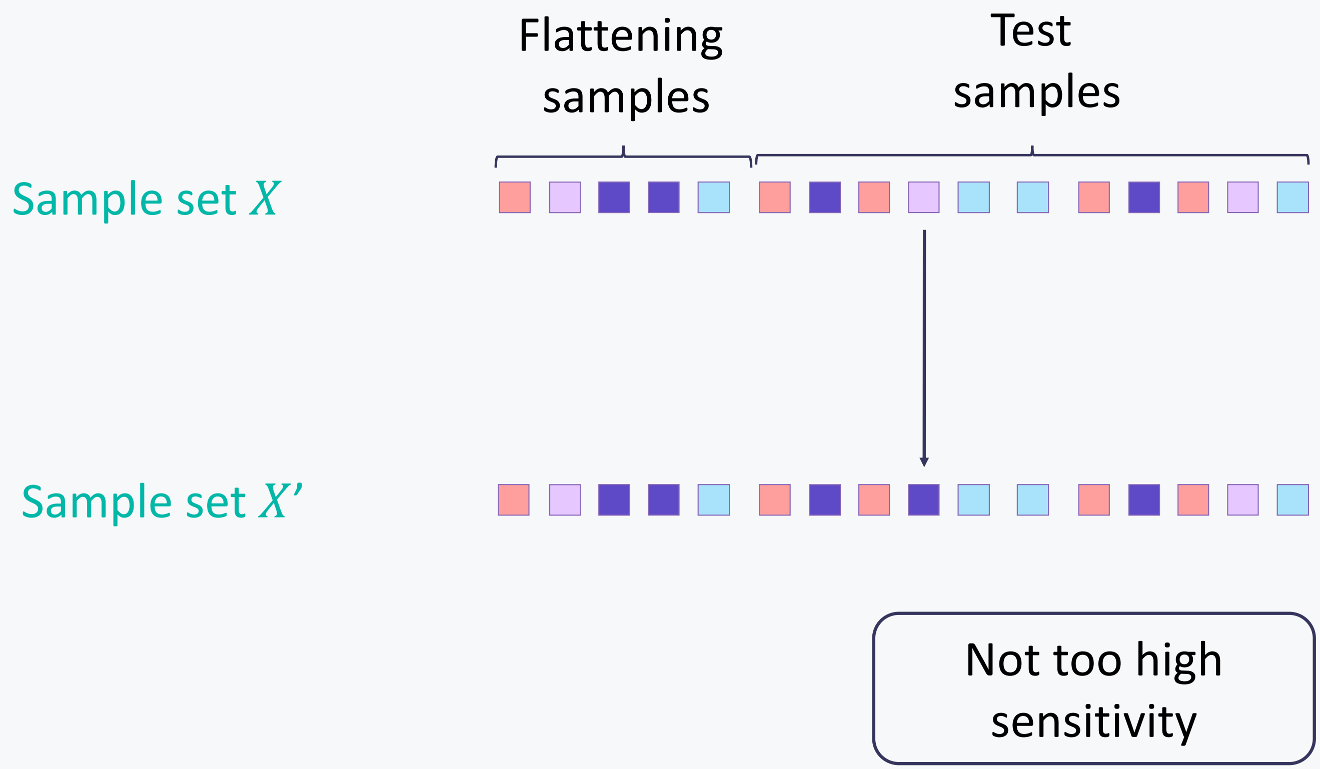


Noise make statistics similar

Noise make statistics similar

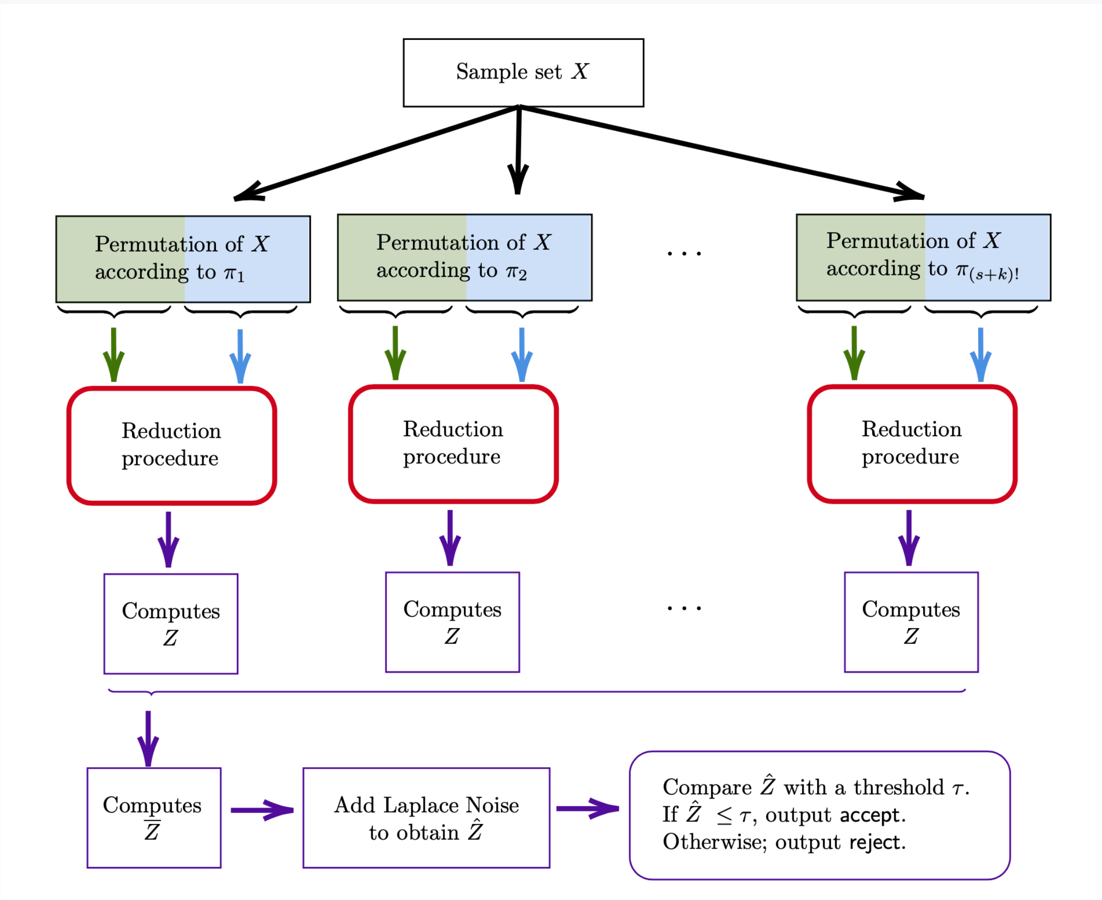


Flattening samples
Test samples
Sample set X
Sample set X'
High sensitivity



Flattening samples
Test samples
Sample set X
Sample set X'
Not too high sensitivity

Our algorithm: derandomization



- Try all partitions for flattening and test samples

- Compute the mean of statistics

New statistic: $\overline{Z} := E_\pi[Z]$

Proof sketch: Why $\overline{Z}$ works

Accuracy

Privacy guarantee

Efficiency: number of samples and time

Proof sketch: Why $\overline{Z}$ works

Accuracy

Privacy guarantee

Efficiency: number of samples and time

- Not independent trials of the algorithms
- Flattening guarantees only worked in average Requires a new analysis

Proof sketch: Why $\overline{Z}$ works

Accuracy	Privacy guarantee	Efficiency: number of samples and time

- Analyze how $\overline{Z}$ changes after changing one sample
- Add noise to hide the change
- Does noise affect accuracy?

Proof sketch: Why $\overline{Z}$ works

Accuracy	Privacy guarantee	Efficiency: number of samples and time

- Exponential time
- Alternative approach with linear time in sample size

Our result on closeness: privacy is almost free!

Theorem

[A, Diakonikolas, Kane, Rubinfeld'19]

There exists a ϵ-private algorithm for testing closeness of two distributions p and q over domain of $[n]$ with error parameter α that uses

$$O\left(\underbrace{\frac{n^{2/3}}{\alpha^{4/3}}+\frac{\sqrt{n}}{\alpha^2}}_{\text{Non-private cost}}+\underbrace{\frac{\sqrt{n}}{\alpha\sqrt{\epsilon}}+\frac{1}{\alpha^2\epsilon}}_{\text{Cost of privacy}}\right)$$

samples from p and q.